Supervised Process of Un-structured Data Analysis for Knowledge Chaining

Matthieu Quantin, Benjamin Hervy, Florent Laroche, Alain Bernard

Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 50 (2016) 436 – 441

www.elsevier.com/locate/procedia

26th CIRP Design Conference

Supervised Process of Un-structured Data Analysis for Knowledge Chaining

Matthieu Quantin[a,b,*], Benjamin Hervy[a], Florent Laroche[a], Alain Bernard[a]

[a] *École Centrale de Nantes, IRCCyN UMR_CNRS_6597, Nantes, France*
[b] *Université de Nantes, CFV EA_1161, Nantes, France*

* Corresponding Author *Tel.:* +33-2-40-37-69-51 *E-mail address:* matthieu.quantin@irccyn.ec-nantes.fr

Abstract

Along the product life-cycle, industrial processes generate massive digital assets containing precious information. Besides structured databases, written reports hold unstructured information hardly exploitable due to the lack of vocabulary and syntax standardization. In this paper we present a methodology and natural language processing approach to exploit these documents. Our method consists in providing connections based on supervised retrieval of domain-specific expressions. No prior document analysis are required to drive the algorithm. It underlines a scale of specificity in pattern visualization. This allows relevant and specific information extraction for feedback (e.g. design stage, after-sales service).

Peer-review under responsibility of the organizing committee of the 26th CIRP Design Conference
Keywords: Knowledge Management; unstructured data; design know-how; semantic information network; Natural Language Processing

1. Introduction

1.1. Context

Knowledge management during industrial processes implies massive digital documents and assets through information systems among the product lifecycle. These documents contain precious information for research and development improvements. Big data techniques are already used for knowledge discovery in order to tackle this issue on structured databases [1]. Besides structured databases of product or process data, there is also unstructured information. An estimate range between 50% and 80% of all the company's data is unstructured [2].

> "There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data". [3]

In this paper we focus on textual analysis of written information rather than massive and structured input data. Therefore, we deal with technical reports, testimonies, and any other written document or collection of documents. Then, we aim to work on knowledge management by chaining knowledge elements rather than data classification and knowledge discovery. The specificity of our work mainly lies on the nature of processed information: contrary to sensors or logs, human actors imply a strong variability in provided information.

1.2. Goals

Over a first phase, the main issue in the process of analyzing unstructured data for knowledge chaining is the domain specificity without any prior information. The process deals with documents that are not conforming to any vocabulary or syntax rules (such as customer reviews or technical notes for example). It underlines a domain-specific vocabulary without labeling or any pre-processing from the document's author. In addition, the process is mainly supervised. Thus, it keeps the results as close as possible from the author work habits and methods.
Over a second phase, our process aims at providing hyper-navigation in corpora based on extracted information. It leads to two main outlines, and a third minor one:

1. Enabling different readings of the same corpus. New levels of reading are based on generic/specific keywords and expressions retrieved by the algorithm.
2. Providing understanding facilities through a located high precision decision-support analytic for unstructured textual corpora. The original text linearity of a single file, or the discontinuity of a corpus is outplayed by networking pieces of extracted information.
3. (minor outline) Offering a massive indexing system on server: anyone could upload a production, that is tagged and linked to other productions or any (other than text) tagged entity.

1.3. Usecase

Based on natural language processing techniques, we aim to demonstrate our proposition through a specific domain: technical history. Our work is inspired by the historians challenges.

Peer-review under responsibility of the organizing committee of the 26th CIRP Design Conference
doi:10.1016/j.procir.2016.04.123

Historical production, definitely human centered, discontinuous and multi-scaled is a perfect experimental field for raw text processing and knowledge chaining. For this article we also used a special corpus, the one from the CIRP paper annals, providing a tangible example for the CIRP community. We will discuss in the last part of this paper the interest of such work for industrial engineering community and how our proposal can be extended to other productions than of the provided use cases.
Thereafter, the term "corpus" will refer to the raw material input: a collection of written documents or a single one, that may contain internal and external references (pictures, quotation, links, etc).

2. Natural language processing for knowledge management

In research and industrial processes as in historical field, only external supports (mainly written) can capitalize the explicit information humans produce. This type of information is unstructured, and differs from any sensor's data output. In a classical conception of the DIKW[1] hierarchy [4], humans cannot produce "pure" data. As History is purely human produced, we have to deal with higher level agency: information to knowledge chaining, not data.

A piece of information is never independent, so the relationship that links to pieces is essential and is part of the knowledge. The aim of our work is to support the importance of these connections. This arise the ethical question, that won't be further discussed here, of the computer influence in (historical) knowledge establishment.

Three assumptions distinguish our work from other Text Mining tools for Knowledge Management:

1. Our work should not depend on any (human or not) data pre-processing step. Raw texts are our only reliable and expectable source of information. We aim to focus on unstructured data, massively produced in every business. This take us away from pure big data issue for decision making as described by the McKinsey Global Institute [5]. It also differs from text analytic oriented projects such as TXM [6] in french, or big data project with clustering for hierarchical structuring [7] in the same field of experiments.
2. Our previous works with historians taught us that information cannot be *a priori* compartmentalized in stringent entities [8]. Unlike biological [9] or manufacturing datamining, no exhaustive data classes exist. Our goal is not to "turn low level data into high-level and useful knowledge" [10], but to process already "compiled" data, i.e. information.
   A framework for text processing in (human) knowledge management should not be first semantically defined. Previously defined topics are at least narrowing and even irrelevant to a text. A POS-tagging or shallow parsing would not allow to match high pattern specificity, nor a core business vocabulary. So we aim to *extract* characteristic entities from a raw text and not to *recognize* already known ones. In this sense we deeply differ from "Named Entity Recognition for Classification" (NERC) processes such as [11] seeking to fill "the lack of publicly available labelled datasets" in french. In the same field of experiments, other works are focused on NERC and neglect the specificity of the input material [12]. Interpretation should be left for the human reading, this work try to stay as close as possible to the text.
3. The black box effect should be avoided. A fully supervised process is necessary for high quality results. We assume that 10 or 20% of mismatch is often not acceptable. Purely manual tagging is a laborious common practice, that hardly reach completeness, but builds high-quality tags. Our supervised process tries to get the best from the machine (completeness) and from the human (high-quality keywords). Moreover, this supervised step balances the machine responsibility (ethical).
   Once again we differ from a classical conception of supervision in NERC: "The main problem with Supervised Learning techniques is that a large amount of tagged data is needed to implement an effective system"[13], because in our case no training set is needed.

Regarding theses existing frameworks and tools, we state that our approach is complementary though significantly different. Complementary with the human "classical" approach of historians, but also with some other text mining tools. This option will be broached as a perspective in section 4.4.

3. Haruspex: a supervised process for knowledge discovery in un-structured data

To achieve our goal of compiling information produced under the form of written documents, we have designed a process based on natural language processing techniques. This process is implemented with Python3 programming language, with a basic GUI for end users. It takes the form of a simple software called Haruspex that takes a corpus as input and produces an undirected graph as output. The output graph is composed of documents or part of documents as nodes (with their metadata) and weighted keywords based relationships as edges. Users can supervise the entire process and have access to internal mechanisms such as the keyword extraction context.

3.1. Overview of the global process

The global process (see fig. 1) is composed of 4 main steps :

1. first step consists in processing input data files : format conversion, files concatenation, splitting, construction of "unit pages" that consists of file or part of file content.
2. second step runs Automatic Natural Acquisition of a Terminology [14] on the previously prepared pages.
3. third step maps the list of keywords from step 2 with their location in the different "pages" from step 1.
4. forth step builds weighted links between the related pages based on tf-idf (term frequency / inverse document frequency) indicator for each keyword.

Those four steps are detailed in the following sections.

[1] Data Information Knowledge Wisdom

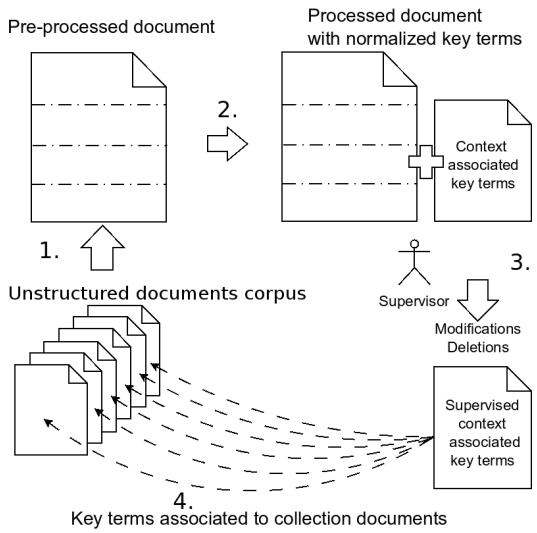

Fig. 1. Simplified view of Haruspex process

3.2. Input and pre-processing

As our program deals with both unique documents and group of documents, the way the corpus is structured depends on the user choices: this step aims to build "pages". A "page" is a meaningful and coherent part of text and the depth of this document structuring is configurable: paragraph, chapter, file, or the concatenation of multiple files. This step is about corpus management.

At the end of this step, we have a text file and an instance of a graph oriented database (using Neo4j). The text file is the concatenation of all the contents with markers to record the original corpus structure. Actually the markers are simple ids written (i.e.: #23_3_6_0) in the concatenated text file to be able to retrieve the section number of the original file (file 23, section 3, subsection 6). The database is a set of nodes, representing the pages by its id, mapped with the original file thanks to the markers. Each node is related to other node types: pictures and external references (bibliography, links, footnotes...) from the original content. There is yet no edge linking the pages among themselves.

Some additional features are provided on demand such as automatic conversion from OpenDocument or pdf, citations and figures management, automatic corpus structuring.

3.3. Automatic Natural Acquisition of a terminology

The concatenated file produced in step 1 is processed here. This step is a new and improved implementation of Natural Language Processing algorithm called ANA[2] [14], dedicated to keywords and expressions extraction without tagging nor training. This algorithm operates on a local level by using syntax and semantic analysis. This is similar to the relation extraction process [15], but does not aim to analyze this relation. The extracted terms are connected to the original corpus thanks to the markers. This step is either highly configurable by the user (thresholds, loops, semantic closeness) or fully automatic with predefined parameters. The output of this step is a list of extracted keywords located in their context. The number of extracted terms is unknown before the end of the process: the main limitation is based on the threshold for occurrences.

[2] Acquisition Naturelle Automatique

3.4. Acquisition post-processing

The output of the previous step (keywords and expressions in their context) is displayed to the supervisor who can remove, modify or add some terms. The user also choose a shape building method for the keywords: shortest shape, most occurring shape, ... This user supervision is critical to keep the control on the content and to reach very high quality extraction (however it may be skipped). A last loop on the text is necessary to spot the new keywords. The output of this step upload the supervised results on the previously created nodes of the database. Each node, representing a page, contains now its own descriptors: the extracted and validated keywords.

3.5. Knowledge chaining

This step consists in creating connections between the nodes (i.e. the pages of step 1) by calculating an value of semantic closeness based on the nodes' keywords. This value is based on the TF-IDF algorithm [16]. The calculated value distinguishes generic links (this connects almost any node to any other) from specific ones (this connect truly some few nodes). All relations between nodes are undirected and have a type: the shared keyword; and a weight: the semantic closeness indicator. Theses links are stored in the database which leads to the construction of a network (see fig.2). So one can request database based on relations properties. This drives us to define the three outlines mentioned in Introduction.

4. Results and perspectives

The quality of the results depends on the way the process is supervised.

On one hand, for general topic spotting the results does not replace a NERC algorithm, on the other hand we extract and build very specific patterns, that no keyword recognition tool provides. Actually, patterns with high relevance and discriminant function are spotted like "*turbine à réducteur á engrenages*" (speed reduction geared turbine), "*chambre syndicale des constructeurs de navires*" (ship builders trade union committee) or "*maçonnerie de moellons hourdés de chaux*"(limewashed rubblestone masonry).

Even if Neo4J database provides the ability to visualize nodes and relations through a web interface, some additional post processing may be rewarding/necessary: the requested content can be dropped on a network visualization software. Using Gephi [17] for example, macro-patterns that could not be observed at a smaller scale are highlighted: lack of information, unlinked sections, or strongly linked sections throughout several original files. Further analysis should be done with experts to discuss these results as new research material.

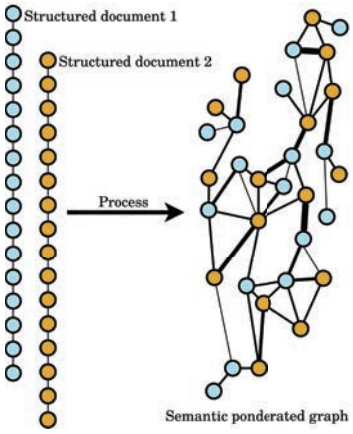

Fig. 2. Networking pieces of semi-structured documents

4.1. Results on a single semi-structured document

We call semi-structured document any text containing basic structure elements like sections and external references. For example this article LaTeX file is a semi-structured document, as many webpages, e-mails, thesis or MS word files. The operating process have been proved on some master thesis, with good results according to the authors.

Under a fully automatic mode, we extract approximately 350 distinct normalized keywords or expressions from a 22k words long corpus with approximately 80% of precision and 98% of recall[3], so the F1 score is around 0.88. That score can be highly increased including the supervision process.

In this case, the main results is to outplay the linearity, displaying a network of related contents. Theses two representations of the knowledge are complementary, and can be combined in a hypertext navigation. As the links are weighted, generic topics are emphasized; as the links are also keyword based, it is possible to follow a narrower topic as a Ariadne's thread.

4.2. Results on a semi-structured corpus

Compared to the previous use case, the extraction quality is mostly the same with additional documents (F1 score: 0.83). The network is changed into a cross navigation. Highlighting the connected parts across the whole corpus is an interesting feature. The most connected content to a section is not necessary the next section of that same document (see fig.2).

In most cases, low-weighted keywords are more generic and used as inter-document (external) links. But some highly specific cross-document connections are also created. This contributes to a serendipity-based reading. An other benefit comes from the identification of lack of information, and the isolated parts of knowledge, weakly or even not connected to other topics at all.

4.3. Results on unstructured corpus

We used Haruspex to process a corpus a compiled pdf documents, these pdf are all papers from the *CIRP Annals - Manufacturing Technology* between beginning of 2008 and end of 2015, we didn't get the internal structure of each pdf file but only the raw text content. This corpus is composed of 109 documents. The date the title and the author of each paper have been recorded in the nodes. Then we extracted around 2300 unique key terms with the expected behavior: very good recall, good precision. Supervision helped to get high quality results. The graph database, binding the pages (articles) among themselves with weighted and keyword based links, enables us to query the corpus. For examples these queries are possible:

- Which are the semantically closest article to a given one?
- How (quantified strength) is evolving the link between virtual and mechanical fields between 2008 and 2015?
- How strong is the link between the medical field, the economical field and the eco-conception field?
- How big is the community speaking about both virtual reality and reverse engineering? Who are these people? What are their non-shared fields of study?

As we are not expert in analyzing this corpus, the aims and the comments of the fig.3 are very basics.
The first graph on the left shows that only one paper mentions "ecodesign" and "economics" (written by Yasushi Umeda), also the papers in the medical field constitute a disjoint set from economics and eco-design.
The second graph in the center shows the hard core of the mechanical field, much more important than the virtual field's one. However many papers mention both of these fields, which are related.
The third graph is about the different streams in the virtual field of studies and their link with the medical one. Biomedical does never refer to the virtual, on the right opposite, the virtual prototyping never concerns the medical field. Also, we notice that one paper is central in this graph, linking medical and virtual fields: the one of S. Ha in 2009.

Note: with large unstructured corpus, the content heterogeneity tends to increase the misconstruction of keywords, matching homographs. Actually, the algorithm matches shapes and syntactic structures, independently of their context of use, nonetheless this can be manually checked by supervision.

On such large document sets, we identify two main benefits from Haruspex:

- the ability to overcome the challenge of indexing large corpus. This is something usual in research projects that involve historical archives analysis and data management plan for interoperability and preservation of data. Therefore, it is often a lot of hard work with few efficiency. Haruspex provides a complementary and semi-automatic indexing with keywords and key terms coming from the documents themselves.
- on a research point of view, Haruspex highlights some extremely specific expressions, thus underlining points of interest for the understanding of large corpus. It provides potential tracks for further research.

As far as knowledge management in design is concerned, Haruspex can bring key ideas from massive sets of documents (users' feedback, interviews, reports and so on). Our process enables the continuous integration of information along the

[3] evaluated by the author of the original corpus and expert of the domain.

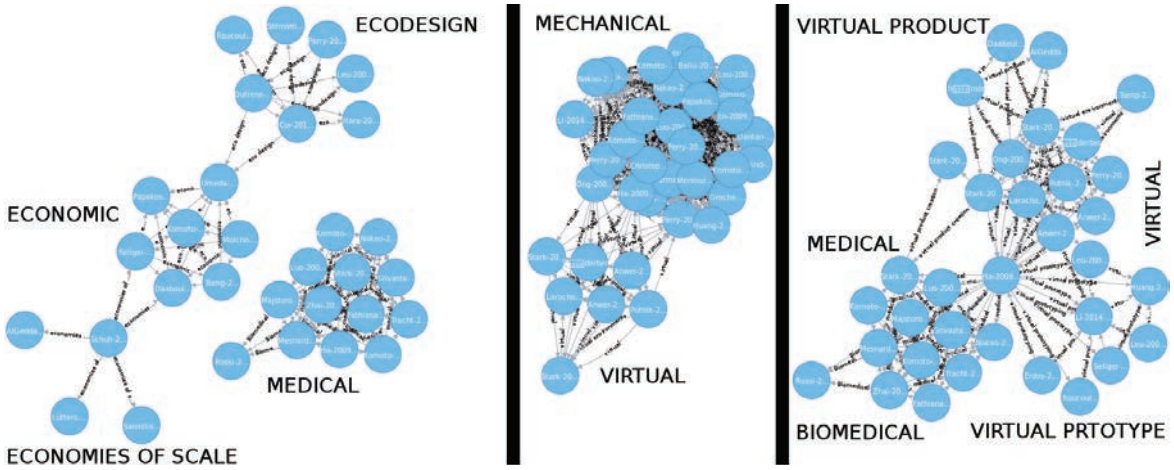

Fig. 3. Graphical representation of the results from querying the database on some relationships and nodes. The links are typed, weighted, despite this is not visible on the graphs above.

product life-cycle. For example with cultural heritage object in museum, we aim to expand its documentation with any new research result or contribution. This paper tends to demonstrate that these benefits can also be valuable for other communities than the heritage one. Our process can be applied on any similar use case, with massive unstructured information available under the format of text documents.

4.4. Perspectives

The process avoid the use of controlled vocabulary, and keep it as a benefit for high precision tagging and corpus analysis. Having a basic controlled vocabulary would be interesting to explore the new documents without starting from scratch. This feature would enable us to extend an existing network with web of data connections. It would lead us to improve our last and still minor outline: the massive indexing system mentioned in Introduction. An external database is queried as reference [18]. This query checks a "normalized shape" for each extracted keywords in DBpedia. This does not overwrite the specific extracted term but provides an other shape for external referencing, and a semantic web compliant output. This step is also supervised. Some more features for the corpus management have to be developed, such as raw metadata management for the nodes, from author (handwritten) or external xml file reading.

This first perspective leads us to differentiating several keyword shapes depending on the use: business intelligence (specific) or massive indexing (standardized). Combining Haruspex with another keyword extracting engines augmented with a supervised layer [19] could be helpful in cases of a highly formalized field (such as medicine).

5. Conclusion

5.1. Benefits

This paper is an attempt to present a supervised process for knowledge chaining based on unstructured data. Our results demonstrate the efficiency of this process. Despite the fact that our algorithm only requires raw text as input, extraction score reach similar results as NERC algorithms. Supervision increases the quality of the algorithm results, avoiding false-positive and giving the ability of final tagging. This feature is appreciated by end users for the final step of knowledge chaining. It ensures compatibility and comprehension between the original text extracted terms and the expert vocabulary. At the end of the process, original texts or text parts are connected based on weighted keywords. The expert has access to different levels of details for further and complementary analysis. This process aims at providing decision aid, thus requiring high quality and controlled terms. Further analysis would suffer from false positive results or a lack of recall.

5.2. Prospect : from knowledge chaining in cultural heritage to un-structured data management in design

As this tool is developed for unstructured and semi-structured text corpus processing without any controlled vocabulary in input, any pre-processing nor dependencies, we could have used it on any subject. Our experiments focus on historical field because we used to work on how technical history issues can help to go beyond industrial complexity. Therefore, we are involved with history research laboratory, providing experts, use cases and feedbacks for our research. Yet, as the design stage information is mainly unstructured [20], we are convinced that our work could be applied to Design with Product Life-Cycle issues in mind. Actually, the importance of such tool for Product Life-cycle Management and called "Product Life-Cyle Analytics" is mentioned in [21] as a reference architecture proposal[4]. Our proposal is in adequation with the unstructured ETL (Extract Transform Load) flow of this architecture. Indeed, we proposed a solution for processing market research documents, patents, internal blogs/wiki, error reports, specifications, feedbacks, memorandum. It would provide connections between all those documents, and give benefit from this amount of hardly reachable information.

[4] ApPLAUDING: An Architecture for Product Lifecycle Analytics with Unstructured Data INteGration.

Acknowledgment

We would like to thanks Eric Lutters, from the University of Twente, for his help and his attentions to our project, and overall for having provided us the CIRP corpus.

References

[1] Hey, T., Tansley, S., Tolle, K.. The fourth paradigm: data-intensive scientific discovery. Microsoft Research; 2009.

[2] Russom, P.. Bi search and text analytics. TDWI Best Practices Report 2007;URL: http://alturl.com/in4mm.

[3] Usama, M.F., Gregory, P.s., Padhraic, S.. From data mining to knowledge discovery in databases. AI Magazine 1996;17:37–54. URL: http://alturl.com/fihaz. doi:10.1145/240455.240463.

[4] Rowley, J.. The wisdom hierarchy: representations of the dikw hierarchy. Journal of Information Science 2007;33(2):163–180. URL: http://alturl.com/7qike. doi:10.1177/0165551506070706.

[5] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. Big data: The next frontier for innovation, competition, and productivity. 2011.

[6] Heiden, S.. The txm platform: Building open-source textual analysis software compatible with the tei encoding scheme. 2010. URL: https://hal.archives-ouvertes.fr/halshs-00549764.

[7] Wang, S., Isaac, A., Charles, V., Koopman, R., Agoropoulou, A., Werf, T.v.d.. Hierarchical structuring of cultural heritage objects within large aggregations. In: Aalberg, T., Papatheodorou, C., Dobreva, M., Tsakonas, G., Farrugia, C.J., editors. Research and Advanced Technology for Digital Libraries; vol. 8092 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg; 2013, p. 247–259. URL: http://link.springer.com/10.1007/978-3-642-40501-3. doi:10.1007/978-3-642-40501-3.

[8] Laroche, F., Bernard, A., Cotte, M.. Knowledge management for industrial heritage. Methods and Tools for Effective Knowledge Life-Cycle-Management 2007;:307–330doi:10.1007/978-3-540-78431-9.

[9] Madeira, S.C., Oliveira, A.L.. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM 2004;1(1):24–45. URL: http://www.ncbi.nlm.nih.gov/pubmed/17048406. doi:10.1109/TCBB.2004.2.

[10] Braha, D.. Data Mining for Design and Manufacturing. Springer; 2002.

[11] Azpeitia, A., Cuadros, M., Rigau, G.. Nerc-fr: Supervised named entity recognition for french. In: Text, Speech and Dialogue; 17th International Conference, TSD. 17; 2014, p. 158–165.

[12] Goerz, G., Scholz, M.. Adaptation of nlp techniques to cultural heritage research and documentation. Journal of Computing and Information Technology 2010;18(4):317. URL: http://alturl.com/ebo4i. doi:10.2498/cit.1001918.

[13] Zelenko, D., Aone, C., Richardella, A.. Kernel methods for relation extraction. The Journal of Machine Learning Research 2003;3:1083–1106. URL: http://dl.acm.org/citation.cfm?id=944919.944964.

[14] Enguehard, C.. Acquisition de terminologie à partir de gros corpus. Informatique & Langue Naturelle 1993;:373–384URL: http://alturl.com/zbmmq.

[15] Chan, Y.S., Roth, D.. Exploiting syntactico-semantic structures for relation extraction. In: 49th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2011, p. 551–560. URL: http://dl.acm.org/citation.cfm?id=2002472.2002542.

[16] Salton, G., Yang, C.S., Yu, C.T.. Contribution to the theory of indexing. Tech. Rep.; 1973. URL: http://bit.ly/1K7s5zY.

[17] Bastian, M., Heymann, S., Jacomy, M.. Gephi: an open source software for exploring and manipulating networks. In: International AAAI Conference on Weblogs and Social Media; vol. 3. 2009, p. 361–362. URL: http://bit.ly/1SVVbop.

[18] Muñoz-García, O., García-Silva, A., Corcho, O., Higuera Hernández, M.d.l., Navarro, C.. Identifying topics in social media posts using dbpedia. In: The NEM summit. Politecnico de Torino; 2011, p. 1–7. URL: http://oa.upm.es/9024/1/Identifying_.pdf.

[19] Turian, J., Ratinov, L., Bengio, Y.. Word representations: a simple and general method for semi-supervised learning. In: 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2010, p. 384–394. URL: http://dl.acm.org/citation.cfm?id=1858681.1858721.

[20] Gero, J.S., Mc Neill, T.. An approach to the analysis of design protocols. Design Studies 1998;19(1):21–61. URL: http://bit.ly/1Q4QlTU. doi:10.1016/S0142-694X(97)00015-X.

[21] Kassner, L., Gröger, C., Mitschang, B., Westkämper, E.. Product life cycle analytics next generation data analytics on structured and unstructured data. CIRP Conference on Intelligent Computation in Manufacturing Engineering 2014;33:35–40. URL: http://goo.gl/LoJg7b. doi:10.1016/j.procir.2015.06.008.